2. Multimedia Data Mining: An Overview

Nilesh Patel and Ishwar Sethi

Summary. Data mining has been traditionally applied to well-structured data. With the explosion of multimedia data methods—videos, audios, images, and Web pages, many researchers have felt the need for data mining methods to deal with unstructured data in recent years. This chapter provides an overview of data mining efforts aimed at multimedia data. We identify examples of pattern discovery models that have been addressed by different researchers and provide an overview of such methods.

2.1 Introduction

Data mining refers to the process of finding interesting patterns in data that are not ordinarily accessible by rudimentary queries and associated results with the objective of using discovered patterns to improve decision making. Traditionally, data mining has been applied to well-structured data, the kind of data that resides in large relational databases. Such data have well-defined, nonambiguous fields that lead easily to mining. In recent years, however, multimedia data—pictures, graphics, animations, audio, videos, and other multimodal sensory streams—have grown at a phenomenal rate and are almost ubiquitous. As a result, not only the methods and tools to organize, manage, and search such data have gained widespread attention but the methods and tools to mine such data have become extremely important too because such tools can facilitate decision making in many situations. For example, the mining of movement patterns of customers from the video routinely collected at shopping malls can be used to improve the layout of merchandize in stores or the layout of shops in the mall.

The mining of multimedia data is more involved than that of traditional business data because multimedia data are *unstructured* by nature. There are no well-defined fields of data with precise and nonambiguous meaning, and the data must be processed to arrive at fields that can provide content information about it. Such processing often leads to nonunique results with several possible interpretations. In fact, multimedia data are often subject to varied interpretations even by human beings. For example, it is not uncommon to have different interpretation of an image by different experts, for example radiologists. Another difficulty in mining of multimedia data are its

heterogeneous nature. The data are often the result of outputs from various kinds of sensor modalities with each modality needing its own way of processing. Yet another distinguishing aspect of multimedia data is its sheer volume. All these characteristics of multimedia data make mining it challenging and interesting.

The goal of this chapter is to survey the existing multimedia data mining methods and their applications. The organization of the chapter is as follows. In Section 2.2, we describe the basic data mining architecture for multimedia data and discuss aspects of data mining that are specific to multimedia data. Section 2.3 provides an overview of representative features used in multimedia data mining. It also discusses the issues of feature fusion. Section 2.4 describes multimedia data mining efforts for concept mining through supervised techniques. Methods for concept mining through clustering are discussed in Section 2.5. Section 2.6 discusses concept mining through the exploitation of contextual information. Event and feature discovery research is addressed in Section 2.7. Finally, a summary of chapter is provided in Section 2.8.

2.2 Multimedia Data Mining Architecture

The typical data mining process consists of several stages and the overall process is inherently interactive and iterative. The main stages of the data mining process are (1) domain understanding; (2) data selection; (3) cleaning and preprocessing; (4) discovering patterns; (5) interpretation; and (6) reporting and using discovered knowledge [1]. The domain understanding stage requires learning how the results of data-mining will be used so as to gather all relevant prior knowledge before mining. Blind application of data-mining techniques without the requisite domain knowledge often leads to the discovery of irrelevant or meaningless patterns. For example, while mining sports video for a particular sport, for example, cricket, it is important to have a good knowledge and understanding of the game to detect interesting strokes used by batsmen.

The data selection stage requires the user to target a database or select a subset of fields or data records to be used for data mining. A proper domain understanding at this stage helps in the identification of useful data. This is the most time consuming stage of the entire data-mining process for business applications; data are never clean and in the form suitable for data mining. For multimedia data mining, this stage is generally not an issue because the data are not in relational form and there are no subsets of fields to choose from.

The next stage in a typical data-mining process is the preprocessing step that involves integrating data from different sources and making choices about representing or coding certain data fields that serve as inputs to the pattern discovery stage. Such representation choices are needed because certain fields may contain data at levels of details not considered suitable for the pattern discovery stage. The preprocessing stage is of considerable importance in multimedia data mining, given the unstructured nature of multimedia data.

The pattern-discovery stage is the heart of the entire data mining process. It is the stage where the hidden patterns and trends in the data are actually uncovered. There

are several approaches to the pattern discovery stage. These include association, classification, clustering, regression, time-series analysis, and visualization. Each of these approaches can be implemented through one of several competing methodologies, such as statistical data analysis, machine learning, neural networks, and pattern recognition. It is because of the use of methodologies from several disciplines that data mining is often viewed as a multidisciplinary field.

The interpretation stage of the data mining process is used to evaluate the quality of discovery and its value to determine whether previous stages should be revisited or not. Proper domain understanding is crucial at this stage to put a value on discovered patterns. The final stage of the data mining process consists of reporting and putting to use the discovered knowledge to generate new actions or products and services or marketing strategies as the case may be. An example of reporting for multimedia data mining is the scout system from IBM [2] in which the mined results are used by coaches to design new moves.

The architecture, shown in Figure 2.1, captures the above stages of data mining in the context of multimedia data. The broken arrows on the left in Figure 2.1 indicate that the process is iterative. The arrows emanating from the domain knowledge block on the right indicate domain knowledge guides in certain stages of the mining process.

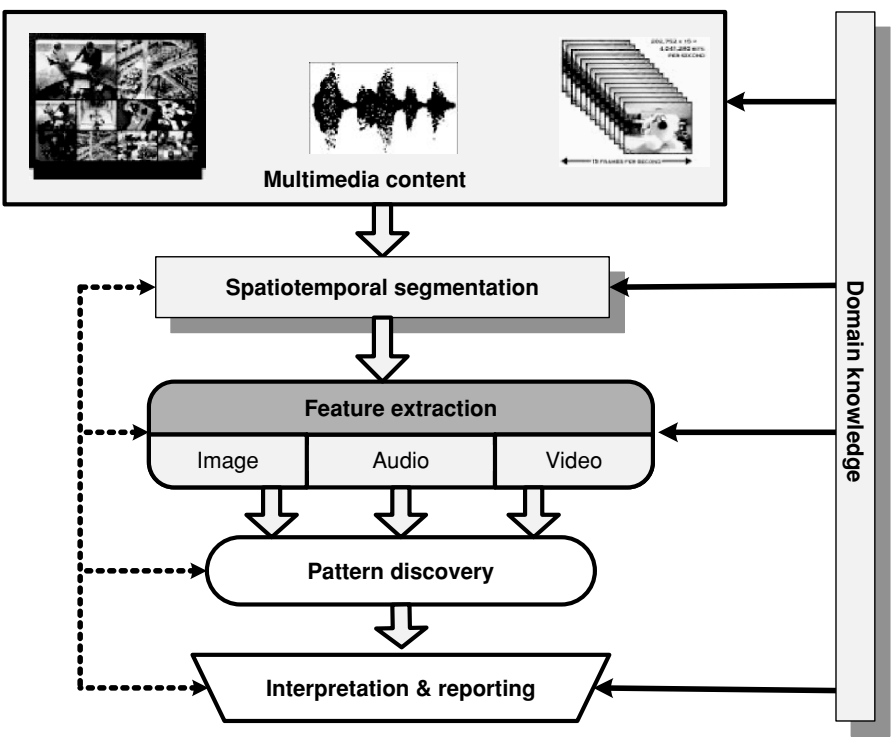

Fig. 2.1. Multimedia data mining architecture.

The spatiotemporal segmentation step in the architecture of Figure 2.1 is necessitated by the unstructured nature of multimedia data. This step breaks multimedia data into parts that can be characterized in terms of certain attributes or features. Thus, in conjunction with the feature extraction step, this step serves the function similar to that of the preprocessing stage in a typical data mining process. In image data mining, the spatiotemporal step simply involves image segmentation. Both region- and edge-based image segmentation methods have been used at this stage in different applications. Although many researchers tend to treat image segmentation for data mining identical to image segmentation needed for computer vision systems, there is an important difference between the requirements for the two segmentations. The image segmentation for a computer vision system should be such that it can operate without any manual intervention and it should be quantitatively accurate so as to allow the vision system to interact with its environment. On the other hand, image segmentation for most data mining applications has no requirement of interacting with its environment. Thus, it can incorporate manual intervention and can be approximate so as to yield features that can reasonably capture the image content. In many image mining applications, therefore, the segmentation step often involves simple blob extraction or image partitioning into fixed size rectangular blocks. With video data, the spatiotemporal step involves breaking the video into coherent collections of frames that can be processed for feature extraction as a single unit. This is typically done via a shot detection algorithm wherein the successive video frames are compared to determine discontinuity along the time axis. A number of shot detection algorithms have been developed in the last 15 years, mostly as a way to organize video data indexing and retrieval [3–8].

Many video mining applications deal with raw or unedited video data, for example, in surveillance and traffic monitoring, as opposed to edited video data typical of entertainment and broadcast video. In such situations, no shot detection type of operation is needed and the unedited video is directly processed to locate events of interests. With audio data, the spatiotemporal step is essentially a temporal step wherein the audio data are segmented either at the phoneme or word level or the data are broken into windows of fixed size.

The pattern discovery step in the multimedia data mining architecture of Figure 2.1 is not much different from mining of traditional and scientific data. Depending on the goal of the discovery stage, the methods for association, classification, clustering, regression, time-series analysis, and visualization are used at this stage. While traditional methods such as decision tree classifier, k-nearest neighbor classifier, k-means clustering, self-organizing feature map (SOFM) continue to be used for pattern discovery in multimedia data, support vector machines (SVMs) have been used widely in recent data mining applications [9, 10]. The SVM is primarily a classifier method that supports both regression and classification tasks and can handle multiple continuous and categorical variables. The SVM finds a solution by mapping the original data to a high-dimensional space where a hyperplane to separate two classes of data can be found. The solution hyperplane is obtained by maximizing the margin of separation between two classes of data in the mapped space. In the standard SVM formulation, the optimal hyperplane is found by solving a quadratic optimization problem. Since

SVM optimization problem grows dramatically with increasing number of training examples, several variations toward using SVMs have been suggested. One such variation is known as *chunking* in which the SVM is trained iteratively with chunks of training cases selected through some scheme. One popular scheme for selecting new training cases is based on *active learning* [11] in which the new training examples are selected through a query function. Such methods have been shown to reduce the overall training time and training set size without sacrificing predictive accuracy.

Examples of pattern discovery models addressed widely in the multimedia data mining community are mining concepts or automatic annotation of multimedia data and discovering events and features. Most multimedia data mining efforts to date have been devoted to the concept mining problem. The reason for such a multitude of efforts lies in the *semantic gap* that users of content-based multimedia information retrieval experience while searching for information through queries such as "find images containing multistory buildings" or "find me a picture of Tajmahal."[1] Dealing with such queries requires providing a retrieval system external knowledge either through manual annotation or through automatic discovering of concepts with or without external supervision. Although manual annotation can be more accurate, it is not practical with ever-increasing multimedia data. Thus, the preferred approach for bridging the semantic gap is to automatically mine concepts or key word associations.

2.3 Representative Features for Mining

Color, edges, shape, and texture are the common image attributes that are used to extract features for mining. Feature extraction based on these attributes may be performed at the global or local level. For example, color histogram of an image may be obtained at a global level or several localized histograms may be used as features to characterize the spatial distribution of color in an image. Similarly, the shape of a segmented region may be represented as a feature vector of Fourier descriptors to capture global shape property of the segmented region or a shape could be described in terms of salient points or segments to provide localized descriptions. There are obvious trade-offs between global and local descriptors. Global descriptors are generally easy to compute, provide a compact representation, and are less prone to segmentation errors. However, such descriptors may fail to uncover subtle patterns or changes in shape because global descriptors tend to integrate the underlying information. Local descriptors, on the other hand, tend to generate more elaborate representation and can yield useful results even when part of the underlying attribute, for example, the shape of a region is occluded, is missing. In the case of video, additional attributes resulting from object and camera motion are used.

In the case of audio, both the temporal and the spectral domain features have been employed. Examples of some of the features used include short-time energy, pause rate, zero-crossing rate, normalized harmonicity, fundamental frequency,

[1] Tajmahal is a famous 15th-century monument in Agra, Uttar Pradesh, India.

frequency spectrum, bandwidth, spectral centroid, spectral roll-off frequency, and band energy ratio. Many researchers have found the cepstral-based features, mel-frequency cepstral coefficients (MFCC), and linear predictive coefficients (LPC), very useful, especially in mining tasks involving speech recognition. While many researchers in this field place considerable emphasis on later processing, Scheirer and Slaney [12] conclude that the topology of the feature space is rather simple, and thus, there is very little difference between the performances of different classifiers. In many cases, the selection of features is actually more critical to the classification performance. In [13], a total of 143 classification features for the problem of general audio data classification are examined to show that cepstral-based features such as the MFCC and LPC provide better classification accuracy than temporal and spectral features.

The MPEG-7 standard provides a good representative set of features for multimedia data. The features are referred as *descriptors* in MPEG-7. The MPEG-7 Visual description tools describe visual data such as images and videos while the Audio description tools account for audio data. A brief description of audiovisual features in the MPEG-7 standard is given here; more details can be found in [14]. The MPEG-7 visual description defines the following main features for color attributes: Color Layout Descriptor, Color Structure Descriptor, Dominant Color Descriptor, and Scalable Color Descriptor. The Color Layout Descriptor is a compact and resolution invariant descriptor that is defined in YCbCr color space to capture the spatial distribution of color over major image regions. The Color Structure Descriptor captures both color content and information about its spatial arrangement using a structuring element that is moved over the image. The Dominant Color Descriptor characterizes an image or an arbitrarily shaped region by a small number of representative colors. The Scalable Color Descriptor is a color histogram in the HSV Color Space encoded by Haar transform to yield a scalable representation. While the above features are defined with respect to an image or its part, the feature Group of Frames-Group of Pictures Color (GoFGoPColor) describes the color histogram aggregated over multiple frames of a video.

The texture descriptors in MPEG-7 are the Edge Histogram Descriptor, Homogenous Texture Descriptor, and Texture Browsing Descriptor. The Edge Histogram Descriptor captures the spatial distribution of edges by dividing the image in 16 nonoverlapping regions. Four directions of edges (0, 45, 90, 135) are detected in addition to nondirectional ones leading to an 80-dimensional vector. The Homogeneous Texture Descriptor captures texture information in a 30-dimensional vector that denotes the energy of 30 spatial-frequency channels computed using Gabor filters. The channels are defined by partitioning the frequency space in angular direction at 30 and by octave division in the radial direction. The Texture Browsing Descriptor is a very compact descriptor that characterizes texture in terms of regularity, coarseness, and directionality.

MPEG-7 provides for two main shape descriptors; others are based on these and additional semantic information. The Region Shape Descriptor describes the shape of a region using Angular Radial Transform (ART). The description is provided in terms of 40 coefficients and is suitable for complex objects consisting of multiple

disconnected regions and for simple objects with or without holes. The Contour Shape Descriptor describes the shape of an object based on its outline. The descriptor uses the curvature scale space representation of the contour.

The motion descriptors in MPEG-7 are defined to cover a broad range of applications. The Motion Activity Descriptor captures the intuitive notion of intensity or pace of action in a video clip. The descriptor provides information for intensity, direction, and spatial and temporal distribution of activity in a video segment. The spatial distribution of activity indicates whether the activity is spatially limited or not. Similarly, the temporal distribution of activity indicates how the level of activity varies over the entire segment. The Camera Motion Descriptor specifies the camera motion types and their quantitative characterization over the entire video segment. The Motion Trajectory Descriptor describes motion trajectory of a moving object based on spatiotemporal localization of trajectory points. The description provided is at a fairly high level as each moving object is indicated by one representative point at any time instant. The Parametric Motion Descriptor describes motion, global and object motion, in a video segment by describing the evolution of arbitrarily shaped regions over time using a two-dimensional geometric transform.

The MPEG-7 Audio standard defines two sets of audio descriptors. The first set is of low-level features, which are meant for a wide range of applications. The descriptors in this set include Silence, Power, Spectrum, and Harmonicity. The Silence Descriptor simply indicates that there is no significant sound in the audio segment. The Power Descriptor measures temporally smoothed instantaneous signal power. The Spectrum Descriptor captures properties such as the audio spectrum envelope, spectrum centroid, spectrum spread, spectrum flatness, and fundamental frequency. The second set of audio descriptors is of high-level features, which are meant for specific applications. The features in this set include Audio Signature, Timbre, and Melody. The Signature Descriptor is designed to generate a unique identifier for identifying audio content. The Timbre Descriptor captures perceptual features of instrument sound. The Melody Descriptor captures monophonic melodic information and is useful for matching of melodies. In addition, the high-level descriptors in MPEG-7 Audio include descriptors for automatic speech recognition, sound classification, and indexing.

A number of studies have been reported in recent literature concerning the performance of MPEG-7 descriptors for a variety of applications. For example, the MPEG-7 shape features have been used to recognize human body posture [15], the defect types in defect images [16], and Zhang and Lu [17] report a comparative study of MPEG-7 shape descriptors and conclude that while both the contour-based curvature scale-space descriptor (CSSD) and the region-based Zernike-moments descriptors perform well for image retrieval, the Fourier descriptors outperform CSSD. In another study, Eidenberger [18] has shown that most MPEG-7 descriptors are highly redundant and sensitive to color shades. Overall, the studies demonstrate that MPEG-7 descriptors are outperformed in several applications by other features. This is not surprising because these descriptors were established to optimize the browsing and retrieval applications of multimedia.

2.3.1 Feature Fusion

An important issue with features extracted from multimedia data is how the features should be integrated for mining and other applications. Most multimedia analysis is usually performed separately on each modality, and the results are brought together at a later stage to arrive at final decision about the input data. This approach is called *late fusion* or *decision-level fusion*. Although this is a simpler approach, we lose valuable information about the multimedia events or objects present in the data because, by processing separately, we discard the inherent associations between different modalities. A series of psychological experiments has shown the importance of synergistic integration of multiple modalities in the human perception system. A typical example of such experiments is the well-known McGurk effect [19]. The other approach for combining features is to represent features from all modalities together as components of a high-dimensional vector for further processing. This approach is known as *early fusion*. The data mining through this approach is known as *cross-modal analysis* because such an approach allows the discovery of semantic associations between different modalities [20].

2.4 Supervised Concept Mining

The concept mining in multimedia is also referred to as automatic annotation or annotation mining. There appears to be three main pattern discovery approaches that have been used for automatic annotation in multimedia data mining. These approaches primarily differ in terms of how external knowledge is provided to mine concepts. In the first approach, an annotator who assigns single or multiple concepts or key words to each multimedia document or its parts provides the external knowledge. This can be viewed as a supervised learning approach. The second approach for automatic annotation is through unsupervised learning or clustering. In this approach, multimedia documents are clustered first and then the resulting clusters are assigned key words by an annotator. Through cluster profiling, rules are next extracted for annotating future documents. The third approach does not rely on manual annotator at all; instead, it tries to mine concepts by looking at the contextual information, for example, the text surrounding an image or the closed caption text of a video. The supervised approach is discussed here; the other two approaches are discussed in the following sections.

Within the supervised framework of automatic annotation, three data mining methods have been used. These are annotation by classification; annotation by association; and annotation by statistical modeling.

2.4.1 Annotation by Classification

Annotation by classification has attracted the most attention, especially for annotating images. The methods for image classification for assigning key words generally differ from traditional object recognition methods in that these methods tend to perform

recognition and classification with little or no segmentation chiefly relying on low-level image features to perform the task. An early example of this is the work of Yu and Wolf [21], who used one-dimensional Hidden-Markov Model (HMM) for classifying images and videos as indoor–outdoor scenes. Some other examples of image classification methods for assigning key words include the works by Vailaya et al. [22], Sethi et al. [23], and Blume and Ballard [24]. Vailaya et al. use a Bayesian framework for classification of outdoor images wherein the images are first divided into the categories of *city* and *landscape* images with landscape images being further subdivided into *sunset, forests*, and *mountain* classes. Their method relies on features derived from color and edge orientation histograms. Sethi, Coman, and Stan [23] use a decision-tree learning scheme to generate classification rules that link the spatial arrangement of colors to predict associated key words. The spatial layout of color in each image is represented by dividing each image into 64 blocks and by calculating the dominant color for each block. The color space used by them is the HSV color space. Their approach generates rules like *If image blocks in the upper half have more than 50% pixels with hue values between 0 and 25, have less than 14% of pixels with hue values between 25 and 41, and have more than 26% of pixels with saturation values between 80 and 100, then the image is considered a sunset image with an estimated accuracy of 90%.*

Several researchers have relied on vector quantization to perform annotation by classification. Blume and Ballard, for example, use a learning vector quantization-based classifier to classify each and every pixel after using Haar wavelet transform to generate a feature vector for every image pixel to capture information about the local brightness, color, and texture. The classified pixels are then grouped into annotated image regions. Blume and Ballard have demonstrated their method to annotate regions with key words such as *sky, forest*, and *water*. Another example of vector-quantization-based classification to predict classification categories is the work done by Mustafa and Sethi [25]. In their approach, a concept-specific codebook is built for each concept using images representing that concept. For example, images representing the concept fire are used to build a codebook for fire. Similarly, images representing *water* are used to build a water-specific codebook. The codebooks are built by using subimages of a certain size. In order to use the codebooks thus built for identifying different concepts in unseen images, every codeword is associated with its own dissimilarity measure that defines whether a particular codeword from a particular concept codebook is sufficiently similar to a block of image pixels in an image being annotated. If a number of codewords from a specific codebook are found similar to image blocks for a given image, then that image is assigned the concept associated with the codebook providing the majority of similar codewords. The approach has been applied to annotate images with three different kinds of *fires, sky, water,* and *grass*, demonstrating that the codebook-based approach is suitable for images whose category can be identified through low-level features such as color. Such features are primary identifiers of what Biederman [26] calls as *mass noun entities* such as *grass, water,* and *snow*. These entities do not have definite boundaries as opposed to count noun entities, for example *airplanes,* with concrete boundaries.

Images with count noun entities cannot be categorized without using shape-based features.

2.4.2 Annotation by Association

Annotation by association methodology is a direct extension of the traditional association rule mining that was developed to mine patterns of associations in transaction databases. Each transaction involves certain items from a set of possible items. Given N transaction and d as the size of the set of possible items, the collection of transactions can be represented as a size $N \times d$ matrix. Since each transaction involves only very few items, the transaction matrix is very sparse. Association mining tries to discover frequent itemsets, that is, the items that appear together frequently in the transaction matrix, in the form of rules wherein the presence of a particular item in a transaction predicts the likely presence of some other items. A typical rule has the form $X \Rightarrow Y$ with support s and confidence c, implying that s% of the transactions contain both X and Y and c% of the transactions that support X also support Y.

Different methods for annotation by association mining differ in terms of how items and transactions are defined to take advantage of existing association rule mining algorithms, for example the well-known Apriori algorithm [27]. An early example of applying association rule mining for image annotation is provided by the work of Ordonez and Omiecinski [28], who consider segmented images to compute the co-occurrences of regions that are deemed similar. The regions are treated as items, and each image constitutes an equivalent of a transaction to extract association rules of the form, "The presence of regions X and Y imply the presence of region Z with support s and confidence c." Ding et al. [29] follow a different strategy to define items and transactions in their work with mining remotely sensed imagery. They divide spectral bands into several small windows, and each window is considered an item. The pixels in their equivalency constitute transactions. Their rule extraction also considers auxiliary information, such as crop yield, at each pixel location to produce rules of the form, "A window in band 1 at $[a_1, b_1]$ and a window in band 2 at $[a_2, b_2]$ results in crop yield y with support s and confidence c." The problem with pixel-level association rules is that pixel-level information is susceptible to noise and furthermore pixels are highly correlated in spatial directions and thus the transactions cannot be considered independent. Tesic et al. [30] present a similar approach to derive the equivalent of the transaction matrix for images. First, the images are partitioned into fixed size rectangular regions. By operating at block level, their method is better at dealing with noise and transaction independence. MPEG-7 textual descriptors are then extracted for each region. A previously constructed learning vector quantizer-based codebook, serving as a visual thesaurus, then provides labels for image regions. The labeled regions are then treated as analogous to items in a transaction database and their first- and second-order spatial co-occurrences are tabulated next. An adaptation of the Apriori algorithm is then used to extract association rules. The method has been used to mine aerial video-graphic images with good success. Another example of annotation by association is the work of Teredesai et al. [31], who use multirelational

rule mining to allow for multiple descriptions for each image arising from multiple sources of labeling.

2.4.3 Annotation by Statistical Modeling

In this approach, a collection of annotated images is used to build models for joint distribution of probabilities that link image features and key words. An early example of this approach is the work of Mori et al. [32], who used a simple co-occurrence model to establish links between words and partitioned image regions. Recently, this approach has started receiving more attention. In the linguistic indexing approach of Li and Wang [33], the two-dimensional multiresolution hidden Markov model (2DMHMM) is used to build a stochastic model for each image category. The 2D MHMM captures statistical properties of feature vectors and their spatial dependence at different levels. The model assumes a first-order Markov chain across the resolutions to define statistical dependencies at different resolutions. The model further assumes that given a block's state, that is, image category label, at any resolution, the corresponding feature vector of the block is conditionally independent of any other states and blocks. This allows the chain rule to be used to compute associated probabilities from training images. Feature extraction is done by partitioning each image into blocks of suitable size, and a six-dimensional feature vector is extracted for each block. The feature vector consists of three components that carry color information of the block and three components that carry texture information. The process of feature extraction is performed at multiple levels for each image giving rise to feature vectors of the same dimensionality that capture information at different levels of details. Li and Wang report modeling of 600 concepts, using an annotated database of 60,000 images. The capability to model such a large number of concepts is one of the strengths of this approach.

Barnard et al. [34] have studied two classes of stochastic models to link images with words. Their approach requires images to be segmented into regions unlike Li and Wang's approach of fixed partitioning. The eight largest regions from each image are selected and a 40-dimensional feature vector, capturing size, shape, color, and texture information, is computed for each region. Figure 2.2 shows one of the models studied by Barnard et al. In this hierarchical model, each node captures relationships between regions and concepts. Higher level nodes capture relationships for key words that are present in many images while nodes at progressively lower levels capture key words that are specific to fewer images. A Gaussian distribution for regions and a multinomial distribution for key words model the joint distribution for probabilities at each node. The results shown by Barnard et al. demonstrate that the models, such as the hierarchical model of Figure 2.2, are able to generate annotation for unseen images; however, the performance depends heavily on the quality of segmentation and manual labeling during training.

Another recent work in this area is due to Jeon et al. [35], who use the relevance-based language models [36, 37] to perform automatic annotation. They assume that every image can be described using a small vocabulary of blobs. The joint distribution of words and blobs is learned using a training set of annotated images to form a model

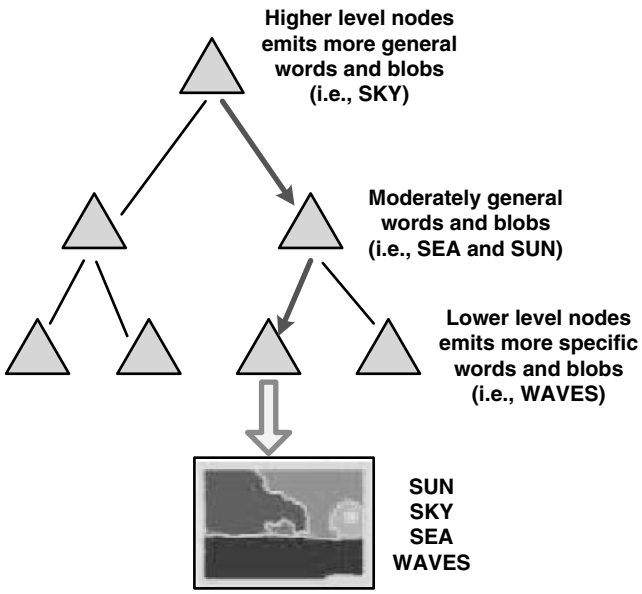

Fig. 2.2. A hierarchical model for joint modeling of regions and key words (adopted from Barnard et al. [34]).

that they call as the cross-media relevance model (CMRM) for images. The CMRM is then used to generate word probabilities for each blob in the test set, which are then combined to build a vector of word probabilities for each image. By keeping the top few probabilities, their approach is then able to generate a few key words for each unseen image.

2.5 Concept Mining Through Clustering

Clustering is another popular data mining methodology that several researchers have used to uncover relationships between key words and images. Clustering-based annotation has been performed at the image level, at the subimage level, and at the region-level after segmentation.

An example of annotation through clustering at the image level is the work of Stan and Sethi [38]. The interesting aspect of this work is cluster profiling in a lower dimensional feature space or in a subspace of the original feature space. Figure 2.3 shows the conceptual architecture of their system to discover relationships between low-level features and key words. Low-level image features such as dominant image colors and their spatial layout are first computed for annotated images. The resulting vectors are next clustered using a hierarchical clustering scheme that allows an easier control over the number of resulting clusters as well as the use of an arbitrary similarity measure. Each of the resulting clusters is analyzed to find the components of the feature

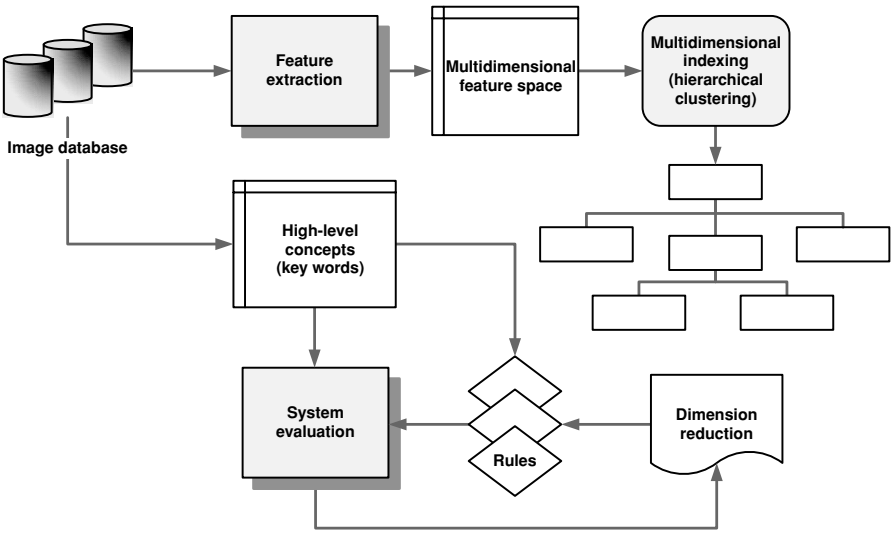

Fig. 2.3. Conceptual architecture for concept mining through clustering (taken from Stan and Sethi [23]).

space that are important for that particular cluster, that is, its own subspace. This is done by ranking features for each cluster in terms of their variance. Features showing variance less than a certain cutoff value are considered important for the corresponding cluster and are retained to specify the subspace for the corresponding cluster. It is this subspace wherein the cluster prototype description is generated. Such a description consists of mean feature values for each subspace feature and the corresponding variances that determine the region of influence for the cluster. Assuming the existence of associated key words for images clustered, key words associated with images in each cluster are next counted and ranked. Only the most frequent key words are then retained to generate mining rules in IF-THEN form. The approach has been applied to generate rules for concepts such as *sunset, landscape*, and *arid*, using color features only. Stan and Sethi [39], in a related work, present another approach for discovering the relationships. In this approach, clusters are visualized through multidimensional scaling in two dimensions. Through visualization, thumbnails of images in a cluster or cluster prototypes are displayed to an annotator who can select a cluster or images close to the cluster prototype for annotation and assign key words as shown in Figure 2.4. Being interactive, the approach offers a compromise between the two extremes of annotation, manual and fully automatic.

While the above examples of work dealt with annotation at image level, doing clustering at subimage level provides better flexibility for annotation because different subimages may represent different concepts. An approach for clustering-based annotation at the subimage level is presented by Mori et al. [32]. In their approach, subimages are clustered through vector quantization to generate codewords. Each image is allowed to be associated with multiple concepts, for example an outdoor image

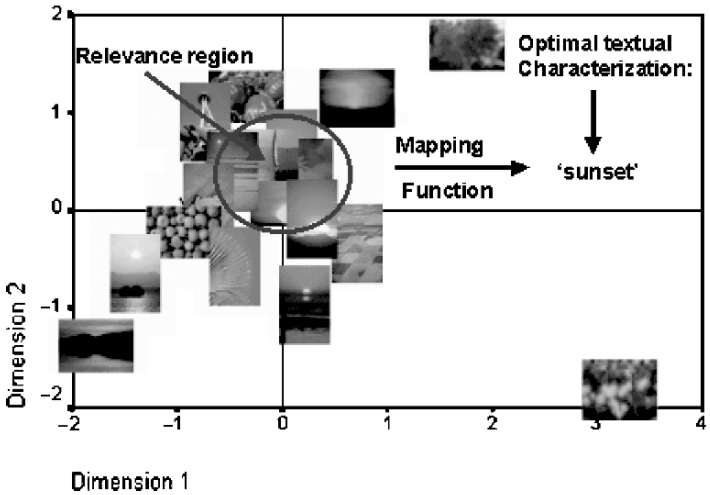

Fig. 2.4. Interactive annotation through multidimensional scaling of cluster centers (taken from Stan and Sethi [39]).

may have *mountains, sky,* and *lake* associated with it. All subimages inherit multiple concepts associated with the parent image. Once the codebook is constructed, inherited concepts for each codeword are used to set up voting probabilities for different concepts given an image. The experimental results presented in [32] indicate that the method is suitable for linking pictures and words; however, the performance could be improved by restricting the domain of pictures and concepts. This is not surprising given that human beings use a great deal of external knowledge in describing the content of an image.

2.6 Concept Mining Using Contextual Information

The text associated with images, for example, in the form of an image caption or the text accompanying an image on a Web page, forms a rich source of external knowledge that can be used to derive semantic links for image features. In a video, the analysis of accompanying audio or closed caption text can yield invaluable information to link visual or audio features with semantic concepts. In all of the examples cited here, contextual information thus available minimizes or rather eliminates the need for an external annotator.

Srihari describes one of the early examples of how text associated with pictures can be used to facilitate image retrieval [40]. Although the work is limited to associating names with faces, it provides an effective example of using external knowledge sources to do annotation. Many methods seeking to link external knowledge sources with images or their features rely on latent semantic indexing (LSI), a popular and powerful technique for text retrieval that has shown to be capable of learning "hidden

associations" between the terms of text documents in the vector model of information retrieval [41, 42]. The LSI approach is based on the singular value decomposition (SVD) method of statistical analysis and is similar to principal component analysis (PCA). The major difference between the two is that the SVD method is applicable when we have a single set of observations and the PCA method applies when we have multiple sets of observations.

The use of LSI for text retrieval involves first forming a term-document matrix. Such a matrix is obtained by first doing preprocessing of the documents to filter out the stop words and convert the remaining words to their root form through stemming. After that, the term weighs for each document are calculated to obtain a vector representation for each document. These vectors then define the term-document matrix, where rows represent the terms and each column represents a document. The singular value decomposition of the term-document matrix then is used to determine a new representation. In the new vector space, each axis corresponds to a "concept" which is related with the terms. It is this relationship that shows similarities between different terms. The previous work [43] in this area has exploited these similarities by applying LSI methodology to bilingual documents, for example in French and English, to determine the associations between French and English words.

There are a few different ways in which LSI has been used to mine concepts in images. In one of the early works involving LSI and image retrieval [44], LSI is used only on text accompanying images to find several concepts that might relate to an image. Although the approach is able to improve retrieval accuracy, no explicit associations between image features and key words are discovered. Consequently, the approach has a limited use. Zhao and Grosky [45] suggest a better approach to exploit LSI. In their approach, the text surrounding an image is processed first to generate a vector representation for the surrounding text. Next, the associated image is processed to extract its low-level features to generate a feature vector representation for the image. A composite vector consisting of appended text and image vectors is then used to represent each image and its surrounding text. These composite vectors then form the columns of the term-document matrix of the LSI approach and the image features and key words of the surrounding text form the terms. The results presented by Zhao and Grosky using 43 different concepts taken from news headlines show that LSI is able to capture relationships between key words and image features.

Instead of using LSI for mining concepts that link key words with visual features, Stan and Sethi [46] have used it to mine associations between different features themselves. Using the bins of a global color histogram as terms, they form the term-document matrix for a collection of images where the global color histogram of each image constitutes its vector representation. By experimenting with 2000 images and a quantized histogram of 192 bins, they have shown that LSI can uncover patterns of color that tend to occur together. These color patterns then can be assigned semantic meanings through supervision.

The other approach to generate concepts for association with images is to perform analysis of the audio and closed captions associated with video. For example, a number of methods for audio classification have been developed that assign audio segments one of the predetermined labels [47–49]. In the case of audio only, the automatic

speech recognition can be performed on segments classified as speech. By analyzing closed captions where available, it is also possible to generate a list of key words that can be associated with a video clip. The key words thus generated from either audio classification or closed captions or both can be used to perform the function of contextual knowledge with LSI to mine concepts for automatic annotation. Kulesh et al. present an approach [50] along these lines in their PERSEUS project where the contextual information from different sources is exploited to track news stories on the Internet for the creation of a personal news portal.

While most work on association mining in multimedia has typically focused on linking multimedia features with key words or captions, recently several researchers have been studying the mining of cues to semantically link different modalities. An example of such a cross-modal association is provided by the work of Li et al. [51] in which three different approaches for cross-modal association are implemented and compared. The three approaches are latent semantic indexing (LSI), cross-modal factor analysis (CFA), and cross-modal canonical correlation analysis (CCA). In the LSI model of cross-modal association discovery, a composite feature vector carrying information from two or more modalities is formed and subjected to singular-value-decomposition to obtain associations. A drawback of this approach as noted by Li et al. is that LSI does not distinguish features from different modalities in the joint space. The set of linear transformations from LSI provide an optimal fit of the overall distribution of different features. However, the optimal solution based on overall distribution may not best represent semantic relationships between features of different modalities, since distribution patterns among features from the same modality will also greatly impact LSI's results. A solution to the above problem is possible by treating features from different modalities as two subsets and focus only on semantic patterns between these two subsets. Under the linear correlation model, the problem becomes one of finding the optimal transformations that can best identify the coupled patterns or the semantic structure between features of two different subsets. Two such transformations are possible through the multivariate statistical analysis methods of factor analysis and canonical correlation analysis. The experimental results presented by Li et al. show that CFA yields the best cross-modal association results. Although CCA follows the same approach of treating different modalities as two subsets, its performance is impacted because associations are more susceptible to noise or small variations in features.

2.7 Events and Feature Discovery

An event in multimedia literature implies an occurrence of an interesting action, for example a car chase. While it is common to characterize an event as an interesting temporal composition of objects, a better characterization of an event is an interesting spatiotemporal instance. This allows us to include spatial combinations of objects, for example the presence of a face in front of a map in an image, as events too. Detection of events has received considerable interest in multimedia literature. For example, there is a large amount of literature related to the detection of events in sports

videos. Methods have been developed to detect and highlight events for basketball [52, 53], baseball [54], soccer [55, 56], and tennis [57]. With the increasing use of cameras for monitoring and surveillance, many researchers have developed methods to detect events in such videos [58, 59]. The primary motivation for large interest in event detection has been that events provide an excellent framework for indexing and summarizing multimedia data.

Event detection implies knowledge of known patterns forming the event; this, in turn, means that specific detectors for different events can be built. The existing literature on event detection exemplifies this where the detectors look for predefined combinations of objects through heuristics, rule-based, or classification methods. Event discovery, on the other hand, implies no prior knowledge of the event characteristics. In fact, the only supposition for event discovery is that the event is something out of the ordinary; that is, an event is an interesting outlier combination of objects. With this viewpoint, the literature on event discovery is sparse and only recently researchers have begun to look for event discovery in multimedia data mining [58, 59].

An example of recent work on event discovery is the work of Divakaran and his group [60, 61] with raw audio and video, which they term as *unscripted media*. Their framework for event discovery is shown in Figure 2.5. In this framework, the multimedia data are windowed and processed to extract features. These features are viewed as time series data, which are converted to a time series of discrete labels through classification and clustering. For example, audio data from a sport broadcast can be classified into a series of distinct labels such as Applause, Cheering, Music, Speech, and Speech with Music. Unusual subsequences in the discrete time series of labels are next detected as outliers by eigenvector analysis of the affinity matrix constructed from estimated statistical models of subsequences of the time series. The length of the subsequences used for statistical model building determines the context in which the events are discovered. The overlap between the subsequences determines the resolution at which the events are discovered. The detected outliers are then ranked on the basis of how well they deviate from background sequences. The interestingness of the ranked outliers is next determined by bringing in the domain knowledge to discover interesting events in the data. Divakaran and his group have successfully applied this methodology to discover events in audio broadcasts of sports and audio captured at traffic intersections. For example, they have shown that their approach was able to discover "ambulance crossing" event in audio data.

Another example of event discovery is the work of Zhong et al. [62], where an unsupervised technique is presented for detecting unusual activity in a large video set. In their method, the video is divided into small segments of equal lengths and each segment is classified into one of the many prototypes using color, texture, and motion features. The detection of unusual events is done by performing a co-occurrence analysis of feature prototypes and video segments. The method has been applied to surveillance video and shown to discover events such as cars making U-turns and backing-off. Another recent example of event discovery work is the semisupervised adapted hidden Markov Model (HMM) framework [17] in which the usual or background event models are first learned from the training data. These models are then used to compute the likelihood of short subsequences of the test data to locate outliers

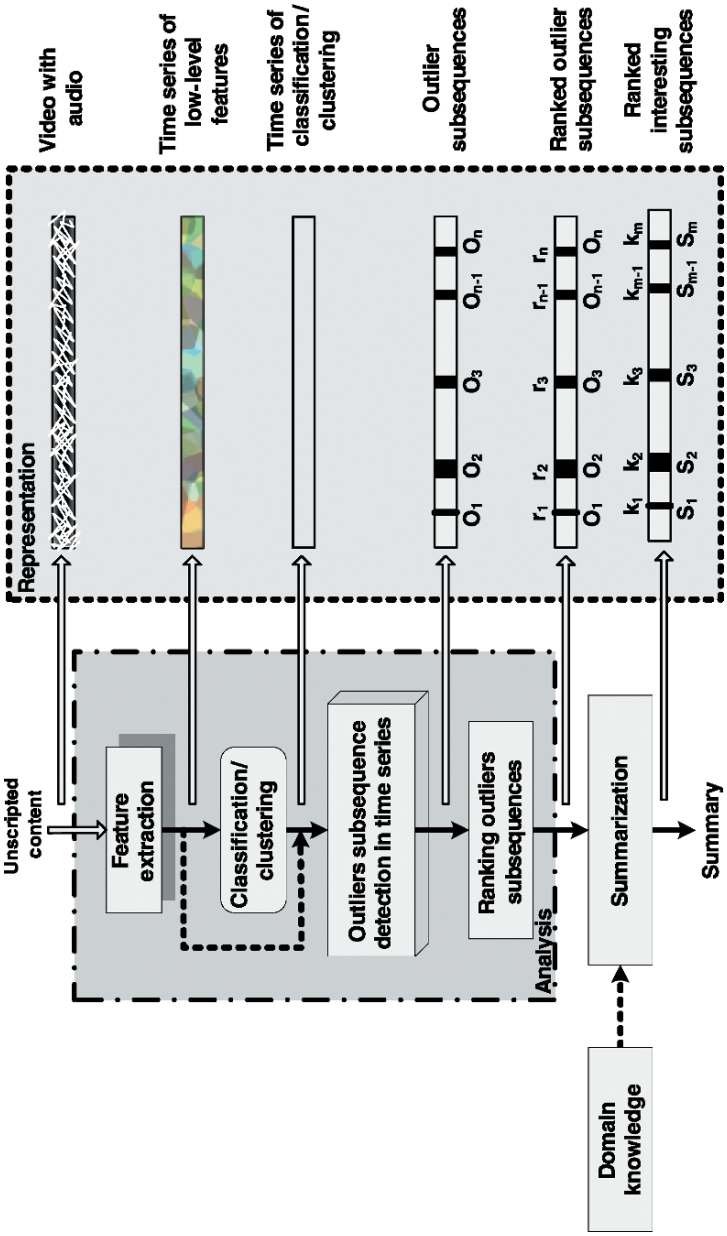

Fig. 2.5. A framework for event detection in unscripted content (adopted from Divakaran et al. [60]).

that show up with small likelihood values. The outliers are then used to adapt the background models to create models for unusual events and locate new unusual event types by iterating the process. The method has been applied to audio and audio-visual data to automatically discover events, such as discovery of interruptions in a multimedia presentation because of questions from audience and laughter from audience.

A related problem to event discovery is the feature discovery problem. A large body of multimedia processing research has dealt with the extraction of features and their use in building indices for content browsing and retrieval. The features described earlier or their variations are typically used for such purposes. While such an approach of using predefined or handcrafted features generally works well, it is tempting to employ data mining methods for the automatic discovery of low-level features that might be best suitable for a given collection of multimedia.

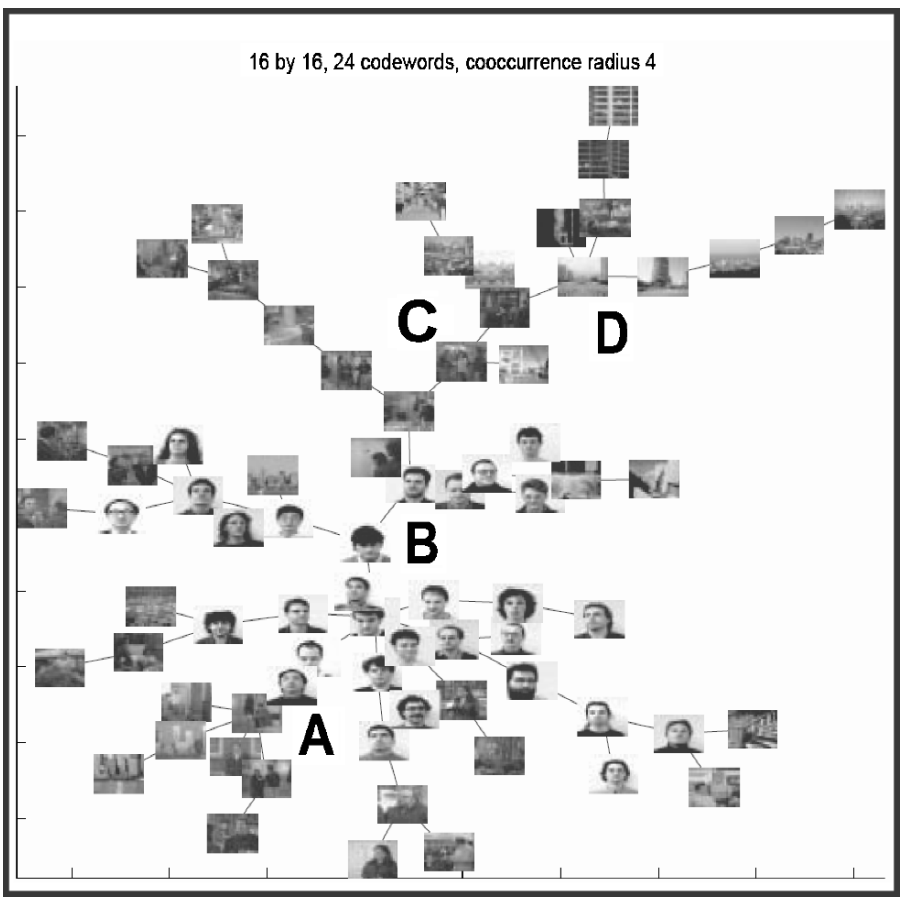

Fig. 2.6. Minimal spanning graph representation of three categories of images using discovered features (taken from Ma et al. [63]).

An example of a feature discovery approach is provided by the work of Mukhopadhyay et al. [63], who have used it to build features for the retrieval and browsing of a collection of images of faces, people, and buildings. Their feature discovery approach is designed for shape features that are extracted from edges of objects in images. After edge extraction, each edge image is partitioned into subimages of a certain size, for example, 16×16. These subimages are then used to construct a codebook whose entries are the low-level features discovered. The salient point of this work is that the codebook is built using the Hausdorff metric [64], which is able to take into account perceptual similarity of edge fragments, an important requirement for retrieval. By looking at the co-occurrences of the discovered features, Mukhopadhyay, Ma, and Sethi have shown that images of different characteristics tend to get clustered in different groups and that there exists a gradual change in clusters as one moves from one cluster to another neighboring cluster as shown in the minimal spanning graph of Figure 2.6.

Lam and Ciesielski [65] describe another example of feature discovery for discovering texture features through genetic programming [66]. In this work, the input is the gray-level histograms of 16×16 subimages. Through genetic programming, various combinations of inputs are searched and evaluated through a fitness function based on how well different texture subimages are separated. The results seem to indicate that discovered texture features are able to yield classification accuracies close to some well-known texture features. Thus, it is a promising approach for building features.

2.8 Conclusion

The multimedia data mining is an active and growing area of research. While the majority of the work has been devoted to the development of data mining methodologies to deal with the specific issues of multimedia data, the origin of the multimedia data mining lies in the pioneering work of Fayyad and his coworkers [67–69] at NASA in the early nineties when they developed several applications including the cataloging of astronomical objects and identification of volcanoes. Since then, several applications of multimedia data mining have been investigated [70–74]. Many of the recent multimedia data mining applications are focused on traffic monitoring and video surveillance, possibly due to increased attention to homeland security. In the coming years, we expect the multimedia data mining applications to grow especially in areas of entertainment and medicine. Almost all of the multimedia data mining efforts to date have been with the centralized data mining algorithms; however, this is expected to change as more and more multimedia content is searched and accessed through peers.

References

1. Sethi I. *Data Mining in Design and Manufacturing*. Kluwer Academic Publishers; 2001.
2. Dimitrova N, Jasinschi R, Agnihotri L, Zimmerman J, McGee T, Li D. *The Video Scout System: Content-Based Analysis and Retrieval for Personal Video Recorders*. CRC Press; 2003.

3. Patel N, Sethi I. Statistical approach to scene change detection. In *Proceedings IS&T/SPIE Conference on Storage and Retrieval for Media Databases*, 1995;2420:329–338.
4. Patel N, Sethi I. *Compressed Video Processing for Cut Detection*. VISP 1996, Vol. 143, pp. 315–323.
5. Lupatini G, Saraceno C, Leonardi R. Scene break detection: A comparison. Research Issues in Data Engineering. In *Workshop on Continuos Media Databases and Applications*, 1998, pp. 34–41.
6. Gargi U, Kasturi R, Strayer S. Performance characterization of video-shot-change detection methods. *IEEE Transaction on Circuits and Systems for Video Technology* 2000;10(1).
7. Lienhart R. Reliable transition detection in videos: A survey and practitioner's guide. *International Journal of Image and Graphics* 2001;1(3):469, 486.
8. Hampapur A, Jain R, Weymouth T. Production model based digital video segmentation. *Multimedia Tools and Applications* 1995;1:9, 45.
9. Joachims T. *Kernel Methods—Support Vector Learning*. MIT Press, 1999.
10. Burges C. A tutorial on support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 1998;2(2):121, 167.
11. Li M, Sethi I. SVM-based classifier design with controlled confidence. In: *Proceedings of 17th International Conference on Pattern Recognition* (ICPR 2004). Cambridge, UK, 2004, pp. 164–167.
12. Scheirer E, Slaney M. Construction and evaluation of a robust multifeature speech/music discriminator. In: *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*, April 1997, pp. 1331–1334.
13. Li D, Sethi I, Dimitrova N, McGee T. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters* 2001;22:533–544.
14. MPEG-7: The generic multimedia content description standard, Part 1. *IEEE MultiMedia* 2002;9(2):78–87.
15. Goldmann L, Karaman M, Sikora T. Human body posture recognition using MPEG-7 descriptors. In: *IS&TSPIE's Electronic Imaging 2004*. San Jose, CA, 2004, pp. 18–22.
16. Pakkanen J, Ilvesmki A, Iivarinen J. Defect image classification and retrieval with MPEG-7 descriptors. In: *Proceedings of the 13th Scandinavian Conference on Image Analysis*, Göteborg, Sweden, 2003, pp. 349–355.
17. Zhang D, Lu G. Evaluation of MPEG-7 shape descriptors against other shape descriptors. *Multimedia System* 2003;9(1):15–30.
18. Eidenberger H. Statistical analysis of MPEG-7 image descriptions. *ACM Multimedia Systems Journal*, 2004;10(2):84–97.
19. McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature* 1976;264:746–748.
20. Li D, Dimitrova N, Li M, Sethi I. Multimedia content processing through cross-modal association. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*. New York: ACM Press, 2003, pp. 604–611.
21. Yu H, Wolf W. Scenic classification methods for image and video databases. In *In SPIE International Conference on Digital Image Storage and Archiving Systems*, Vol. 2606, 1995, pp. 363–371.
22. Vailaya A, Jain A, Zhang H. On image classification: City vs. landscape. *Pattern Recognition* 1998;31:1921–1936.
23. Sethi I, Coman I, Stan D. Mining association rules between low-level image features and high-level semantic concepts. In: *Proceedings SPIE Conference on Data Mining and Knowledge Discovery*, April 2001.

24. Blume M, Ballard D. Image annotation based on learning vector quantization and localized Haar wavelet transform features; 1997. Available from: citeseer.ist.psu.edu/blume97image.html
25. Mustafa A, Sethi I. Creating agents for locating images of specific categories. In: *IS&T Electronic Imaging 2004*, San Jose, CA, 2004.
26. Biederman I. Recognition by components: A theory of human understanding. *Psychological Review* 1987;94:115–147.
27. Agrawal R. *Fast Discovery of Association Rules*. Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press 1996, pp. 307–328.
28. Ordonez C, Omiecinski E. Discovering association rules based on image content. In: *ADL '99: Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*. Washington, DC: IEEE Computer Society; 1999, p. 38.
29. Ding Q, Ding Q, Perrizo W. Association rule mining on remotely sensed images using p-trees. In: *In Proceedings of PAKDD*, 2002.
30. Tesic J, Newsam S, Manjunath B. Mining image datasets using perceptual association rules. In: *SIAM International Conference on Data Mining*, Workshop on Mining Scientific and Engineering Datasets. San Francisco, CA; 2003, pp. 71–77.
31. Teredesai A, Ahmad M, Kanodia J, Gaborski R. CoMMA: A framework for integrated multimedia mining using multi-relational associations. *Knowledge and Information Systems: An International Journal*, in press.
32. Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: *MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
33. Li J, Wang J. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2003;25(9):1075–1088.
34. Barnard K, Duygulu P, Forsyth D. Trends and advances in content-based image and video retrieval. In Press.
35. Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: *In Proceedings of the 26th international ACM SIGIR Conference*, 2003, pp. 119–126.
36. Lavrenko V, Choquette M, Croft W. Cross-lingual relevance models. In: *Proceedings of the 25th Annual International ACM SIGIR Conference*, 2002, pp. 175–182.
37. Lavrenko V, Croft W. Relevance-based language models. In: *In Proceedings of the 24th International ACM SIGIR Conference*, 2001, pp. 120–127.
38. Stan D, Sethi I. Mapping low-level image features to semantic concepts. In: *Proceedings IS&T/SPIE Conference on Storage and Retrieval for Media Databases*, San Jose, CA, 2001.
39. Stan D, Sethi I. eID: A system for exploration of image databases. *Information Processing and Management* 2003;39:335–361.
40. Srihari R. Automatic indexing and content-based retrieval of captioned images. In: *IEEE Computer*, Vol. 28, 1995, pp. 49–56.
41. Berry M, Dumais S, OíBrien G. Using linear algebra for intelligent information retrieval. *SIAM Review* 1995;37:573–595.
42. Deerwester S, Dumai S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. Journal of American Society for Information Science 1990;41:391–407.
43. Dumais S, Landauer T, Littman M. Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing. In: *Proceedings SIGIR*, pp. 16–23.

44. Cascia M, Sethi S, Sclaroff S. Combining textual and visual cues for content-based image retrieval on the world wide web. In: *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, 1998, pp. 24–28.
45. Zhao R, Grosky W. *Distributed Multimedia Databases: Techniques and Applications*. Hershey, PA: Idea Group Publishing.
46. Stan D, Sethi I. Color patterns for pictorial content description. In: *Proceedings of the 17th ACM Symposium on Applied Computing*, 2002, pp. 693–698.
47. Wijesekera D, Barbara D. Mining cinematic knowledge: Work in progress. In: *Proceedings of International Workshop on Multimedia Data Mining (MDM/KDD'2000)*, Boston, MA, 2000, pp. 98–103.
48. Snoek C, Worring M. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications* 2005;25(1):5–35.
49. Lau R, Seneff S. Providing sublexical constraints for word spotting within the ANGIE framework. In: *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 263–266.
50. Kulesh V, Petrushin V, Sethi I. Video Clip Recognition Using Joint Audio-Visual Processing Model. In: *ICPR (2002)*, 2002, pp. 500–503.
51. Li D, Li M, Nevenka D, Sethi I. Multimedia content processing through Cross-Modality Association. In: *Proceedings of the 11th ACM Int'l Conf. Multimedia*, Berkeley, CA, 2003, pp. 604–611.
52. Zhou W, Vellaikal A, Jay-Kuo C. Rule-based video classification system for basketball video indexing. In: *ACM Multimedia Workshops*, 2000, pp. 213–216.
53. Nepal S, Srinivasan U, Reynolds G. Automatic detection of "Goal" segments in basketball videos. In: *MULTIMEDIA '01: Proceedings of the Ninth ACM International Conference on Multimedia*. New York: ACM Press, 2001, pp. 261–269.
54. Rui Y, Gupta A, Acero A. Automatically extracting highlights for TV baseball programs. In: *MULTIMEDIA '00: Proceedings of the Eighth ACM International Conference on Multimedia*. New York: ACM Press, 2000, pp. 105–115.
55. Gong Y, Sin L, Chuan C, Zhang H, Sakauchi M. Automatic parsing of TV soccer programs. In: *IEEE Conference on Multimedia Computing and Systems*, 1995.
56. Tovinkere V, Qian R. Detecting semantic events in soccer games: Towards a complete solution. In: *Proceedings of ICME 2001*, Tokyo, Japan, 2001.
57. Wang J, Parameswaran N. Analyzing tennis tactics from broadcasting tennis video clips. In: *11th International Multimedia Modelling Conference (MMM'05)*, pp. 102–106.
58. Tucakov V, Ng R. Identifying unusual spatiotemporal trajectories from surveillance videos. In: *In Proceedings of 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98)*, Seattle, WA, 1998.
59. Oh J, Lee J, Kote S, Bandi B. Multimedia data mining framework for raw video sequences. In: Zaiane SJSimoff, Djeraba Ch, editors. Mining multimedia and complex data, Lecture Notes in Artificial Intelligence, Vol. 2797, Springer, 2003, pp. 18–35.
60. Divakaran A, Miyaraha K, Peker K, Radhakrishnan R, Xion Z. Video mining using combinations of unsupervised and supervised learning techniques. In: *SPIE Conference on Storage and Retrieval for Multimedia Databases*, Vol. 5307, 2004, pp. 235–243.
61. Goh S, Miyahara K, Radhakrishan R, Xiong Z, Divakaran A. Audio-visual event detection based on mining of semantic audio-visual labels. In: *SPIE Conference on Storage and Retrieval for Multimedia Databases*, Vol. 5307, 2004, pp. 292–299.
62. Zhong H, Shi J, Visontai M. Detecting unusual activity in video. In: *Proc. CVPR*, 2004.
63. Mukhopadhyay R, Ma A, Sethi I. Pathfinder networks for content based image retrieval based on automated shape feature discovery. In: *Sixth IEEE International Symposium on Multimedia Software Engineering (ISMSE 2004)*, FL, 2004.

64. Ma A, Mukhopadhyay R, Sethi I. Hausdorff metric based vector auantization of binary images. In: *Proceedings Intíl Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, 2003, pp. 315–320.
65. Lam B, Ciesielski V. Discovery of human-competitive image texture feature extraction programs using genetic programming. *GECCO* 2004;2:1114–1125.
66. John K. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: The MIT Press, 1992.
67. Smyth P, Burl M, Fayyad U, Perona P. Knowledge discovery in large image databases: Dealing with uncertainties in ground truth. In: *In Proc. of AAAI-94 Workshop on KDD*, Seattle, WA, 1994, pp. 109–120.
68. Fayyad U, Weir N, Djorgovski S. Automated analysis of a large-scale sky survey: The SKICAT System. In: *In Proceedings 1993 Knowledge Discovery in Databases Workshop*, Washington, DC, 1993, pp. 1–13.
69. Fayyad U, Smyth P. Image database exploration: Progress and challenges. In: *In Proceedings 1993 Knowledge Discovery in Databases Workshop*, Washington, DC, 1993, pp. 14–27.
70. Zhu X, Wu X, Elmagarmid A, Feng Z, Wu L. Video data mining: Semantic Indexing and event detection from the association perspective. *IEEE Transactions on Knowledge and Data Engineering* 2005;17(5):665–677.
71. Yoneyama A, Yeh C, Jay-Kuo C. Robust vehicle and traffic information extraction for highway surveillance. *EURASIP Journal on Applied Signal Processing*. 2005;14:2305–2321.
72. Yoneyama A, Yeh C, Jay-Kuo C. Robust traffic event extraction via content understanding for high way surveillance system. *IEEE International Conference on Multimedia and Expo*, 2004.
73. Za O, Han J, Li Z, Hou J. Mining multimedia data. In: *CASCON '98: Proceedings of the 1998 Conference of the Centre for Advanced Studies on Collaborative Research*. IBM Press, 1998.
74. Oh J, Lee J, Kote S. Real Time Video Data Mining for Surveillance Video Streams. In: *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Seoul, Korea, 2003, pp. 222–233.